Text Summarization For Indian Languages

Abstract

Indian languages have a rich linguistic diversity and are spoken by a large population, yet automatic text summarization for these languages has received little attention in academic research. This gap is particularly challenging due to the scarcity of large, publicly available datasets, which is in stark contrast to languages like English and Chinese. In this work, we tackle the challenges highlighted by the Indian Language Summarization Task (ILSUM) by developing tailored solutions for Indian English, Bengali, Hindi, and Gujarati. Our proposed approach combines Named Entity Recognition (NER) with advanced multilingual transformer models to generate summaries that are semantically rich. We also take into account the complex nature of Indian languages, such as code-mixing and script-mixing. To further expand the applicability of our model, we incorporate Optical Character Recognition (OCR) capabilities, enabling the processing of non-digital data formats. Using the datasets provided by ILSUM 2023, we aim to address common issues and enhance Indian language processing through systematic analysis and algorithmic improvements.

1 Introduction

In the fast progressing field of natural language processing (NLP), text summarization is an important job that serves many purposes. As digital content is growing at a very high rate in different languages, there exists an urgent requirement for efficient techniques to summarize it while taking into account specific linguistic characteristics and difficulties. This paper is centered on creating a summarization model especially for Hindi, which happens to be one among the most spoken languages globally.

Hindi, being a language, gives particular difficulties for text summarization. Hindi has complex mix of code and script where English phrases are integrated into the Hindi text especially in news articles. This creates problem for current summarization models as they might not correctly understand the meaning of mixed-language content.

Additionally, the Hindi language has a lot of digital content available in it. This includes things like news, literature and social media posts. Having such variety makes automated summarizing tools very important as they can quickly summarize big amounts of text into smaller but still meaningful summaries. The usual methods for text summarization may not work well with Hindi because of its special language characteristics and cultural subtleties.

Preprint. Under review.

2 Motivation

The research was driven by the requirement to tackle the difficulties of Hindi text summarization. In Hindi text, there is a lot of code-mixing and script mixing which makes it hard for current techniques to summarize effectively. This makes it more necessary to have specific summarization models that can handle the details of Hindi language correctly.

In addition, the rise of Hindi digital content on different platforms shows why it is necessary to have automated summarization tools that can quickly and effectively sort through large amounts of information. Be it news articles, social media postings or discussions in online forums - being able to create short summaries in Hindi language benefits users as well as developers.

We have observed that there is a scarcity of summarization models available for Hindi language. This creates a barrier in using natural language processing techniques on Hindi documents, as summarization is an important part of many applications such as document understanding and information retrieval. By creating a dedicated summarization model for Hindi, we intend to overcome this obstacle and offer a useful resource for numerous applications like information finding, content suggestion and learning languages. Our study supports the general objective of improving NLP methods for languages that are not well represented, promoting fairness and ease of use in the time we live where everything is digitalized.

Our work is a thorough study of text summarization methods that are made for the unique traits of Hindi language. The main aim is to help in quick and correct summarizing of Hindi text across different areas.

3 Literature Survey

The project has been documented and studied well in the past. We came across several pieces of literature published on this topic, which are covered below. These works highlight the challenges and opportunities in automatic text summarization for Indian languages. Verma and Om (2020) (2) and Baruah et al. (2019) (3)provide a comprehensive overview of the state-of-the-art: discussing various techniques such as extractive and abstractive summarization, graph-based methods, and rule-based approaches. They emphasize the need for language-specific resources, such as stemmers, stop-word lists, and named entity recognizers, to improve the quality of the input data. Kumar and Yadav (2015) (4) propose an extractive summarization approach for Hindi based on thematic term selection. Their architecture involves preprocessing steps like stop-word removal and stemming, followed by sentence scoring based on the presence of thematic terms and cue phrases. The sentence scoring formula used is:

$$S_{c_j} = \frac{\sum_i M[i,j]}{|Terms|}$$

where Sc_j is the score of sentence j, M[i, j] is the value of the cell [i, j] in the document-term matrix, and |Terms| is the total number of terms in the document. The authors also incorporate linguistic knowledge from Hindi WordNet for post-processing and removing extraneous phrases. Urlana et al. (2022) (1) presented a study using pre-trained sequence-to-sequence models for English, Hindi, and Gujarati text summarization. They experiment with state-of-the-art architectures like MT5, mBARTLARGE 50, and IndicBART and investigate the impact of k-fold cross-validation and data filtering techniques on model performance. The proposed architecture involves preprocessing, fine-tuning, and evaluation steps, focusing on adapting pre-trained models to the summarization task using language-specific datasets.

Scope for improvement in these approaches includes developing unified frameworks that combine statistical and linguistic features, exploring the use of deep learning architectures like transformers or sequence-to-sequence models, creating large-scale, high-quality datasets for training and evaluation, and addressing challenges specific to Indian language text data, such as code-mixing and script-mixing.

4 Methodology

4.1 Dataset Overview 1 4 3

The dataset provided by ILSUM includes headings, articles, and summaries, aimed at generating fixed-length extractive or abstractive summaries for each news article. These articles are sourced from various leading newspapers in India, comprising over 15,000 unique and relevant articles for each language. One notable feature of this dataset is the presence of code and script mixing, which involves multilingual headlines, such as Hindi titles containing English phrases. This reflects a realistic scenario where textual content isn't confined to a single script or language but often comprises an amalgamation of two or more languages.

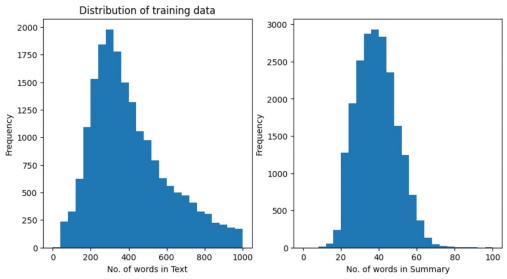

Figure 1: Overview of Training Data

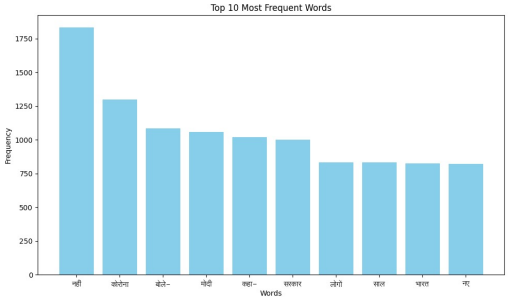

Figure 2: Frequency of words

For Named Entity Recognition, the data subset utilized comprises of 21,657 rows, each with tokenized text and corresponding numerical tags for various named entity categories such as festivals, games, languages, literature, and others. The tagging follows a 'B-' prefix for the beginning and 'I-' for the interior of named entities, with 'O' representing non-entity tokens. The columns include 'tokens' for the Hindi words or phrases and 'ner_tags' for their entity classifications. The dataset is structured into 70% training, 15% validation, and 15% test portions, facilitating the training of a BiLSTMCRF model that leverages both past and future context for accurate entity identification. This resource is available on Hugging Face, provided by HiNERoriginal and we have leveraged the test portion for our training.

4.2 Proposed Architecture

Our model introduces a novel approach by integrating several advanced functionalities that go beyond mere text summarization. In addition to generating concise summaries of Hindi text, it is equipped with a Named Entity Recognition (NER) component. The integration of NER within the model is fundamental to enhancing the semantic richness of the summaries produced. The goal is to systematically identify and categorize key entities such as names, organizations, and locations, where the NER component augments the textual data, emphasizing key elements that would contribute to the contextual relevance of the output. This methodology ensures that critical information is present in the summaries, thereby retaining the important details in the condensed text.

Additonally, Optical Character Recognition (OCR) allows our model to digitize and interpret text from scanned documents, images, and other similar media. This diversifies the input sources for the model. OCR capability is also critical for processing non-digital data formats, and it effectively broadens the application horizon of the proposed model. By including documents that presently exist in physical form, there will be a seamless transition from analog to digital text processing. This multi-faceted approach enables a more versatile application, catering to a broader range of input formats and delivering richer, more informative output.

Figure 3: Wordcloud

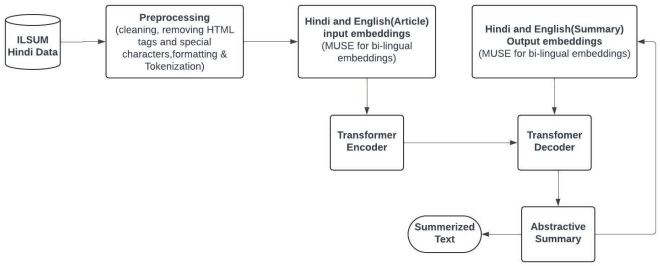

Figure 4: Proposed Architecture with fastText (Futue Work)

4.3 Model Fine-tuning

To achieve state-of-the-art performance in Hindi text summarization, we have fine-tuned two pre-trained language models: IndicBART and mT5-base. Fine-tuning these models allows us to leverage their pre-existing knowledge and adapt them specifically to the task of Hindi summarization.

IndicBART is a multilingual BART (Bidirectional and Auto-Regressive Transformers) model that has been pre-trained on a vast corpus of Indian languages, including Hindi. By fine-tuning IndicBART on a dataset of Hindi text and their corresponding summaries, we enable the model to learn the intricacies and patterns specific to Hindi summarization. The fine-tuning process involves training the model on a large-scale Hindi summarization dataset, where the model learns to generate concise and coherent summaries by minimizing the difference between the generated summaries and the ground-truth summaries.

Similarly, we fine-tune the mT5-base model, which is a multilingual variant of the T5 (Text-to-Text Transfer Transformer) model. mT5-base has been pre-trained on a massive corpus of text from various languages, including Hindi, using a self-supervised learning approach. By fine-tuning mT5-base on the Hindi summarization dataset, we adapt the model's language understanding and generation capabilities to the specific task of Hindi summarization.

5 Results

The presented results show the performance of different models on the tasks of Hindi text summarization [1], for the first 100 summaries from the validation dataset, and Named Entity Recognition (NER) [2].

For Hindi text summarization, the Google mT5 base model outperforms the IndicBart model across various evaluation metrics. The mT5 base model achieves higher scores in ROUGE-1, ROUGE-2, ROUGE-L, and BERT Score F1, indicating its superior ability to generate summaries that closely match the reference summaries. The higher cosine similarity score of mT5 base also suggests that it captures the similarity between the generated summaries and the original text more effectively than IndicBart.

The mT5 base model was trained on a larger dataset compared to IndicBart. This difference in training data size may contribute to the performance gap between the two models. The mT5 base model's exposure to a more extensive and diverse set of Hindi text samples likely enhances its ability to generate accurate and coherent summaries.

In the NER task, the model demonstrates promising results. With an accuracy of 95.79%, the NER model accurately identifies and classifies named entities in the majority of cases. The model performs well in identifying and classifying named entities across all 23 tags present in the dataset.

Metric	Google mT5 base	IndicBart
Dataset Size	21,123	9,857
Trained Samples	21,123	7,885
Average ROUGE-1	**58.88**	28.82
Average ROUGE-2	**56.95**	23.89
Average ROUGE-L	**57.36**	27.46
Average BERT Score F1	**78.95**	76.88
Average Cosine Similarity	**43.10**	37.90

Table 1: Results for Hindi Text Summarization

Metric	Value
Dataset Size	21,657
Number of Tags	23
Trained Samples	15,160
F1 Score (Macro)	**63.05%**
Accuracy	**95.79%**

Table 2: Results for Named Entity Recognition (NER)

6 Observations

For the combined tasks of summarization and NER, we encountered challenges in task synchronization due to the disparate nature of the datasets employed. The NER model's limited exposure to the vocabulary present in the iLSUM dataset reduced its effectiveness in identifying terms for the articles used in summarisation. Expanding the NER model's dataset to encompass a broader vocabulary and a greater diversity of tags could potentially refine its classification capabilities and improve overall performance.

In the summarisation task, ROUGE measures the similarity between the generated summary and the reference summaries based on n-gram overlap and word sequences. BERT score, on the other hand, utilizes contextualized word embeddings from BERTto compute the similarity between the generated summary and the reference summaries. Unlike ROUGE, BERT score takes into account the context of words in the summaries, capturing nuances of meaning more effectively. Our BERT score is better than ROUGE score for the Hindi language summarization task implying that the generated summaries are coherent. This suggests that the model is effectively capturing the semantic meaning and context of the text.

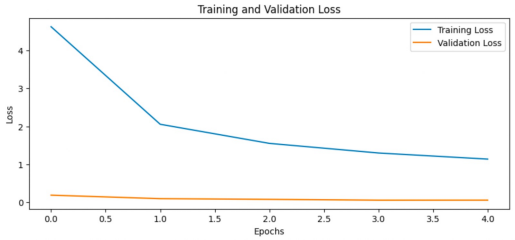

Figure 5: Loss Plot for NER Task

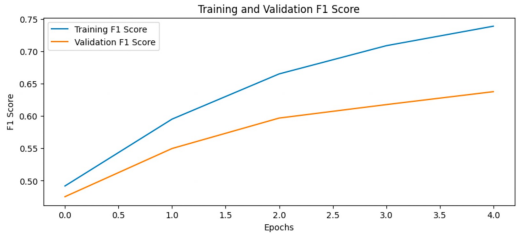

Figure 6: F1 score plot for NER Task

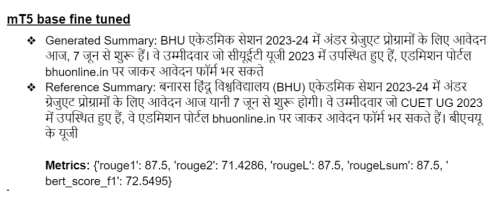

Figure 7: mT5 base generated summary

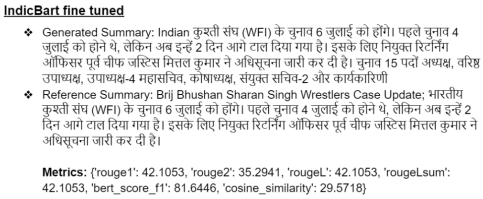

Figure 8: Indic-bart generated summary

7 Conclusion and Future Work

In our approach to model training for summarization, access to more robust computing resources would allow the model to accommodate longer summary outputs and a greater number of training epochs. Furthermore the integration of pre-trained models such as mT5 large can be considered to further enhance performance. In addition, it is evident from the convergence graph that there is still room for improvement in the NER task, suggesting the potential for further training beyond the current five epochs.

For our future work a model from the ground up can be developed, rather than fine-tuning existing state-of-the-art models. Our suggested approach involves using MUSE embeddings to achieve bilingual embedding for English and Hindi, addressing the challenges of code-switching and script variation. As incorporating FastText embeddings can enhance the semantic richness of summaries. We propose a multi-headed transformer architecture with an encoder-decoder mechanism to create abstractive summaries. This architecture is scalable, potentially adaptable to larger datasets and other Indian languages.

For innovative expansion, the NER model, once trained on a broader dataset, can not only be employed to extract key terms but also be used to assign accurate tags. In extending the application of NLP, this task can be further extended to translate the produced summaries into multiple languages, broadening accessibility while maintaining data integrity. This step will ensure the summaries reach a more diverse audience with minimal data distortion.

8 References

Demo Video Link:
https://drive.google.com/file/d/1jC_Tz2vQ-hlBy3rpXiCiTGqw-FoaJmMO/view?usp=drive_link

References

[1] Urlana, Ashok Bhatt, Sahil Surange, Nirmal Shrivastava, Manish. (2023). Indian Language Summarization using Pretrained Sequence-to-Sequence Models.

[2] Pradeepika Verma, Hari Om, MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization, Expert Systems with Applications, Volume 120, 2019, Pages 43-56, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2018.11.022.

[3] Baruah, Nomi Sarma, Shikhar Borkotokey, Surajit. (2019). Text Summarization in Indian Languages: A Critical Review. 1-6. 10.1109/ICACCP.2019.8882968.

[4] Kumar, K.V., Yadav, D.: An ımprovised extractive approach to Hindi text summarization. In: Informatiove System Design Intelligence Applications, pp. 291–300. Sprınger, New Delhi (2015)

[5] Tangsali, Rahul Pingle, Aabha Vyawahare, Aditya Joshi, Isha Joshi, Raviraj. (2022). Implementing Deep Learning-Based Approaches for Article Summarization in Indian Languages. 10.48550/arXiv.2212.05702.

[6] Namrata Kumari and Pardeep Singh. 2023. Hindi Text Summarization Using Sequence to Sequence Neural Network. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 22, 10, Article 239 (October 2023), 18 pages. https://doi.org/10.1145/3624013